

Лекція 5

МЕТОДИ ОБРОБЛЕННЯ ТА АНАЛІЗУ ЕКСПЕРИМЕНТАЛЬНИХ ДАНИХ

5.1. Методика оброблення масиву даних

Дуже часто мають місце випадки, коли вибіркова сукупність даних містить дані, значини яких сильно (на перший погляд) відрізняються від інших. У зв'язки з цим деякі дослідники приймають суб'єктивне і необґрунтоване рішення щодо видалення «сумнівних» даних із статистичного обробітку.

Водночас, бракувати ті чи інші дані можна лише за умови наявності прямих доказів того, що вони є результатом грубої помилки при їх отриманні.

Одним із механізмів такого аналізу є загальновідомий критерій Ірвіна (λ). Дійсну його значину (λ_d) знаходять із виразу:

$$\lambda_d = (X_n - X_{n-1})/\sigma,$$

- де X_n – максимальна «сумнівна» значина масиву даних;
 X_{n-1} – значина із масиву даних, найближча до максимальної;
 σ – середнє квадратичне відхилення масиву даних.

Табличну значину критерію Ірвіна (λ_T) визначають наступним чином:

- для довірчої ймовірності 90%:

$$\lambda_T = \frac{2}{\sqrt{N}} + 0,60,$$

де N - кількість даних вибірки (масиву);

- для довірчої ймовірності 95%:

$$\lambda_T = \frac{2,5}{\sqrt{N}} + 0,75;$$

- для довірчої ймовірності 99%:

$$\lambda_T = \frac{3}{\sqrt{N}} + 1,15.$$

«Сумнівну» величину масиву X_n вибраковують тільки тоді, коли виявиться, що $\lambda_d > \lambda_T$. Аналогічно перевіряють і «сумнівну» найменшу величину масиву експериментальних даних.

Наприклад, нехай нами отримано десять ($N = 10$) вимірів якого-небудь параметру: 5; 6; 5; 7; **8**; 6; 5; 6; 7; **12**.

Сумнів викликає остання значина масиву – «12». Найближчою до неї є п'ята значина параметру – «8». Звідси $X_n = 12$, а $X_{n-1} = 8$.

Розрахувавши для даного масиву середнє квадратичне відхилення σ , знаходимо, що $\lambda_d = 1,89$. Разом з цим, для довірчої ймовірності 95% $\lambda_T = 1,54$. Оскільки в цьому випадку $\lambda_d > \lambda_T$, то вимір масиву, значина якого дорівнює 12, є сумнівним і його слід вибракувати.

Натомість, у аналогічному масиві даних 5; 6; 5; 7; **8**; 6; 5; 6; 7; **10** вимір зі значиною 10 не може бути вибракуваним, оскільки тут $\lambda_d = 1,26 < \lambda_T = 1,54$.

Більшість коливних явищ у сільськогосподарському виробництві представляють собою випадкові стаціонарні і ергодичні процеси. Механізми їх обробітку нині практично повністю формалізовані і при використанні комп'ютерів з програмним забезпеченням не викликають будь-яких труднощів.

Нестационарність того чи іншого випадкового процесу може бути за математичним очікуванням, дисперсією і за тим та другим одночасно.

Коли має місце нестационарність за дисперсією, то, за рекомендацією проф. А.Б. Лур'є, нестационарну ділянку слід просто видалити (рис. 5.1), а решту даних обробляти як єдину генеральну сукупність.

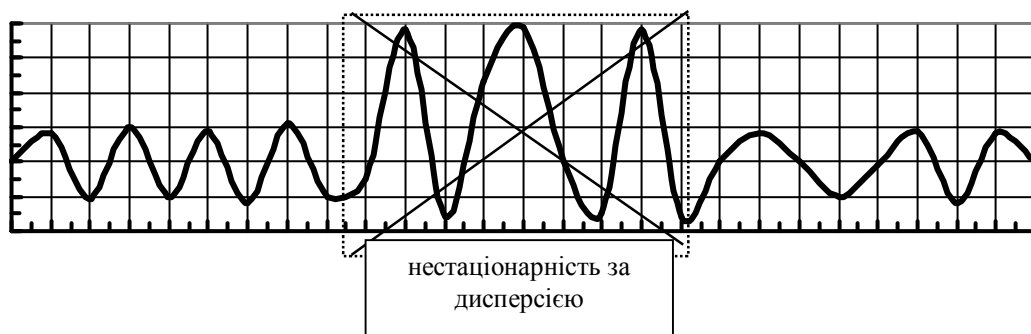


Рис. 5.1– Вид випадкового процесу з нестационарністю за дисперсією

Значно частіше той чи інший випадковий процес може бути нестационарним (причому лінійно) за математичним очікуванням.

Нехай випадковий процес зміни того чи іншого параметру описується кривою $f-f$ (рис. 5.2).

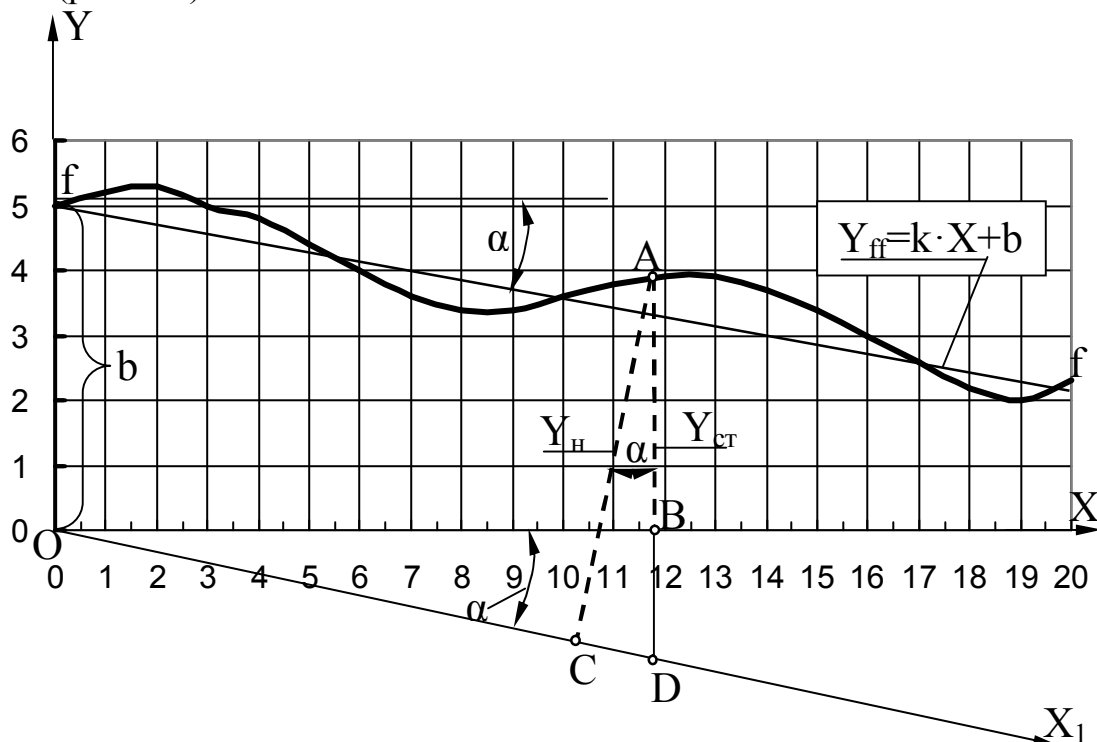


Рис. 5.2 – Схема обробки процесу, нестационарного за математичним очікуванням

На нестационарність за математичним очікуванням цього коливного процесу вказує той факт, що пряма лінія $Y_{ff} = k \cdot X + b$, яка апроксимує криву **f-f**, нахилена до осі OX під кутом α . В результаті, зі збільшенням значини регресора (X) ордината досліджуваного процесу ($Y_{ст}$), замість коливання відносно апроксимаційної прямої Y_{ff} , поступово зменшується.

У принципі, для досягнення поставленої мети криву **f-f** слід повернути проти ходу годинникової стрілки на той же кут α . Здійснити це можна наступним чином. Із початку системи координат YOX треба провести пряму OX_1 , паралельну лінії Y_{ff} (див. рис. 5.2). Далі, стару ординату ($Y_{ст}$) кожної точки кривої **f-f** (т. А, наприклад) слід замінити на нову ($Y_{н}$), отриману відносно лінії OX_1 .

Із рис. 5.2 випливає, що $Y_{н} = AD \cdot \cos \alpha$.

У свою чергу

$$AD = Y_{ст} + BD.$$

Враховуючи, що

$$k = \operatorname{tg} \alpha,$$

можемо записати:

$$BD = -k \cdot X.$$

З урахуванням вищевикладеного остаточно отримуємо:

$$Y_{н} = (Y_{ст} - k \cdot X) \cdot \cos[\operatorname{arctg}(k)]. \quad (5.1)$$

Залежність (5.1) дає можливість ординати лінійно нестационарного коливного процесу перетворити у стаціонарний за математичним очікуванням масив.

Практичне застосування викладеної вище методики продемонструємо на наступному прикладі. В процесі експериментальних досліджень було отримано масив ординат відхилення траєкторії руху просапного МТА від прямої лінії. Виявилося, що цей процес (крива 1, рис. 5.3) є нестационарним за математичним очікуванням, оскільки апроксимаційна пряма $y = 0,0035 \cdot x + 5,7238$ не паралельна до базової лінії (осі OX).

Після оброблення експериментальних даних з урахуванням залежності (5.1) отримано новий масив ординат, який представляє практично стаціонарний процес коливань траєкторії руху просапного МТА (крива 2, рис. 5.3).

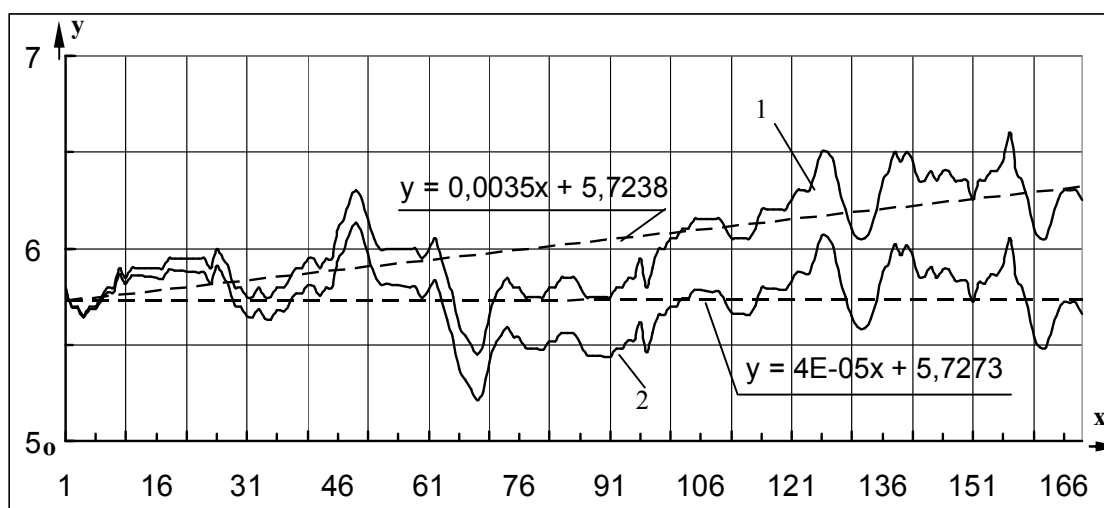


Рис. 5.3 – Нестационарний (1) та стаціонарний (2) за математичним очікуванням випадові процеси коливань траєкторії руху просапного МТА

5.2. Розрахунок та аналіз основних статистичних характеристик

5.2.1 Середня арифметична значина

Методика розрахунку середньої арифметичної значини того чи іншого масиву даних на сучасному етапі абсолютно формалізована і не викликає будь-яких ускладнень. Окремо стоїть питання правильного використання цієї статистичної кількісної характеристики. Особливо при порівнянні двох статистично випадкових процесів.

На практиці може виявитися, що кількісно мала різниця між середніми арифметичними є не випадковою. В такому випадку говорять, що нуль-гіпотеза про рівність порівнюваних статистичних характеристик на тому чи іншому прийнятому статистичному рівні значущості **відхиляється**.

Часто буває навпаки: відносно значна кількісна різниця між середніми значинами виявляється випадковою і нуль-гіпотеза про їх рівність **не відхиляється**. В цьому випадку обидві статистичні характеристики репрезентують одну і ту ж генеральну сукупність (масив) експериментальних даних.

Одним із способів оцінювання суттєвості різниці між середніми арифметичними значинами двох **незалежних** і однакових за об'ємом вибірок даних (масивів) є **t**-критерій. Розраховується він наступним чином:

$$t = \frac{X_1 - X_2}{\sqrt{S_{x1}^2 + S_{x2}^2}}, \quad (5.2)$$

де X_1, X_2 – середні арифметичні значини двох масивів даних;

S_{x1}, S_{x2} – похибки порівнюваних середніх арифметичних X_1 і X_2 .

В загальному випадку величини S_{x1} і S_{x2} знаходять із виразів:

$$S_{x1} = \sigma_1 / \sqrt{N_1};$$

$$S_{x2} = \sigma_2 / \sqrt{N_2},$$

де σ_1, σ_2 – середні квадратичні відхилення (стандарти) масивів N_1 і N_2 .

Оскільки в даному випадку $N_1 = N_2 = N$, то

$$S_{x1} = \sigma_1 / \sqrt{N};$$

$$S_{x2} = \sigma_2 / \sqrt{N},$$

і залежність (5.2) можна представити у такому варіанті:

$$t = \frac{(X_1 - X_2) \cdot \sqrt{N}}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \quad (5.3)$$

Якщо виявиться, що $t \geq t_{\text{теор}}^1$, то нуль-гіпотеза щодо відсутності суттєвої різниці між середніми арифметичними X_1 і X_2 відхиляється. За таких обставин вважають, що статистична характеристика X_1 не випадково більша за статистичну характеристику X_2 .

В протилежному випадку (коли $t < t_{\text{теор}}$) різниця між X_1 і X_2 знаходиться в межах випадкових коливань на відповідному статистичному рівні значущості. А такий факт принаймні не заперечує, що між оцінюваними середніми арифметичними значинами не існує різниці.

¹ таблична величина

Розглянемо наступний приклад. Нехай треба оцінити середні значини глибини оранки двох процесів, кожен із яких представлений 100 вимірюваннями даного оцінюваного параметру, тобто $N = 100$. Розрахунками встановлено, що

$$X_1 = 26,5 \text{ см}; \quad X_2 = 25,3 \text{ см}; \quad \sigma_1 = \pm 1,3 \text{ см}; \quad \sigma_2 = \pm 1,5 \text{ см}.$$

Підставивши ці дані у формулу (5.3), отримуємо, що $t = 6,05$. Теоретична ж (тобто таблична) значина цього критерію для статистичного рівня значущості 0,05 $t_{\text{теор}} = 1,98$. Оскільки $t = 6,05 > t_{\text{теор}} = 1,98$, то нуль-гіпотеза щодо відсутності суттєвої різниці між середніми арифметичними X_1 і X_2 відхиляється. Тобто, середня значина глибини оранки $X_1 = 26,5$ см не випадково більша за характеристику $X_2 = 25,3$ см. Слід зазначити, що X_1 не випадково більша за X_2 і для статистичного рівня значущості 0,001 (тобто для ймовірності 99,9%), оскільки навіть у цьому випадку $t = 6,05 > t_{\text{теор}} = 3,39$.

Якби у обох процесів за тих же X_1 та X_2 середні квадратичні відхилення були більшими і становили:

$$\sigma_1 = \pm 3,5 \text{ см}; \quad \sigma_2 = \pm 3,5 \text{ см},$$

то тоді б із виразу (5.3) отримали, що $t = 2,42$ і на статистичному рівні значущості 0,05 нуль-гіпотеза про рівність порівнюваних середніх значин знову відхиляється, тому що $t = 2,42 > t_{\text{теор}} = 1,98$.

Водночас, на статистичному рівні значущості 0,01 нуль-гіпотеза про рівність порівнюваних середніх значин уже не відхиляється, оскільки $t = 2,42 < t_{\text{теор}} = 2,63$. Тобто із довірчою імовірністю 99% можна стверджувати, що різниця між середніми значинами глибини оранки $X_1 = 26,5$ см і $X_2 = 25,3$ см є статистично випадковою і вони обидві репрезентують одну і ту ж генеральну сукупність (тобто один і той же випадковий процес). На даному статистичному рівні значущості у дослідника немає підстав стверджувати, що величина X_1 більша за X_2 , хоча арифметично це дійсно так.

Досліднику треба чітко знати до чого приводить спроба зменшення об'єму масиву вимірюваних даних N . Так, за умов

$$X_1 = 26,5 \text{ см}; \quad X_2 = 25,3 \text{ см}; \quad \sigma_1 = \pm 1,3 \text{ см}; \quad \sigma_2 = \pm 1,5 \text{ см},$$

але $N = 10$, тому отримуємо $t = 1,91$.

А це означає, що навіть на статистичному рівні 0,05 між середніми значинами глибини оранки $X_1 = 26,5$ см та $X_2 = 25,3$ см немає істотної різниці, оскільки $t = 1,91 < t_{\text{теор}} = 1,98$. І це в той час, коли за умови $N = 100$ ця ж сама різниця між оцінюваними параметрами була істотною як на статистичному рівні значущості 0,05, так і на рівні значущості 0,001.

Перевірити нуль-гіпотезу можна і іншим способом. Для цього використовують найменшу істотну різницю (НІР) між порівнюваними середніми арифметичними. Якщо дійсна різниця між X_1 і X_2 є більшою за НІР або дорівнює їй, тобто $X_1 - X_2 = d \geq \text{НІР}$, то величина d признається значущою і нуль-гіпотеза про рівність порівнюваних характеристик на відповідному статистичному рівні значущості відхиляється. І навпаки, коли $d < \text{НІР}$, то величина є несуттєвою і середні арифметична X_1 є закономірно більшою за X_2 .

Для знаходження величини НІР спочатку визначають похибку різниці середніх арифметичних s_d : $s_d = \sqrt{S_{x1}^2 + S_{x2}^2}$.

Наступним кроком розраховують значини найменших істотних різниць на відповідному статистичному рівні значущості. Наприклад,

$$НІР_{05} = t_{05} \cdot s_d - \text{для довірчої ймовірності } 95\%;$$

$$НІР_{01} = t_{01} \cdot s_d - \text{для довірчої ймовірності } 99\%.$$

У цих формулах t_{05} , t_{01} – теоретичні (табличні) значини t- критерію для статистичних рівнів значущості 0,05 і 0,01 відповідно.

Із викладеного вище бачимо, що на величину різниці між порівнюваними середніми арифметичними суттєво впливає похибка їх оцінювання (S_x). Із теорії математичної статистики відомо, що в цілому похибка середньої арифметичної тим менша, чим менше варіює досліджуваний процес (менше величина σ) і чим більша кількість вимірювань його ординат (N).

Науковці досить часто оперують статистичною оцінкою, яка виражена у відсотках від відповідної середньої арифметичної (X) і називається відносною похибкою вибіркової середньої ($S_{x\%}$):

$$S_{x\%} = (S_x \cdot 100) / X.$$

В деяких літературних джерелах величину $S_{x\%}$ іноді називають «точністю досліду». При цьому вважають, що коли $S_{x\%}$ перевищує 5%, то слід удосконалювати методику досліду, а якщо $S_{x\%} > 7...8\%$, то результати досліджень взагалі не можна вважати достовірними.

Водночас, існує досить обґрунтована думка, яку повністю поділяє і автор даного підручника, що таке визначення відносної похибки середньої значини не відповідає дійсності, а тому не пропонується до ужитку.

Наприклад, при дослідженні глибини загортання насіння пшениці сівалкою були отримані дві середні значини X : 5 і 6 см відповідно. Похибка оцінювання цих характеристик виявилась однаковою і рівною $S_x = 0,25$ см. В результаті для першого варіанту маємо $S_{x\%} = (0,25 \cdot 100) / 5 = 5\%$, а для другого $S_{x\%} = (0,25 \cdot 100) / 6 = 4,2\%$. Виникає питання, на якій підставі «точність досліду» у другому варіанті слід вважати вищою, ніж у першому?

5.2.2 Дисперсія та середнє квадратичне відхилення

Дані статистичні характеристики є основними мірками варіації (коливання, розсіювання) вивчаемого випадкового процесу.

В загальному випадку для визначення дисперсії (D) статистично випадкового процесу не потрібно знати значину середньої арифметичної. Процедура розрахунку цієї характеристики передбачає знання величини та кількості ординат вибірки.

Дисперсія є завжди позитивною, а її розмірність дорівнює квадрату розмірності досліджуваного параметра. Оскільки це не завжди зручно, то застосовується характеристика, яка дорівнює кореню квадратовому із дисперсії. Цілком зрозуміло, що вона має розмірність досліджуваного параметра і називається середнім квадратичним відхиленням або стандартом (σ). Значина стандарту завжди записується із знаком « $\pm$ ». Наприклад, $\sigma = \pm 3,5$ см; $\sigma = \pm 25$ кН; $\sigma = \pm 1,5$ кг тощо.

Для порівняння дисперсій двох процесів D_1 і D_2 з об'ємами вибірок N_1 і N_2

використовують загальновідомий F- критерій Фішера:

$$F = D_1/D_2,$$

де між оцінюваними дисперсіями має виконуватись співвідношення $D_1 > D_2$.

Табличну значину F-критерію (F_T) знаходять в залежності від вибраного статистичного рівня значущості і числа незалежних вимірювань f_1 і f_2 :

$$f_1 = N_1 - 1;$$

$$f_2 = N_2 - 1.$$

Якщо розрахована значина F більша або дорівнює F_T , то з вибраною довірчою ймовірністю можна стверджувати, що дисперсія D_1 більша за дисперсію D_2 . Інакше між цими статистичними характеристиками суттєвої різниці немає і нуль-гіпотеза про їх рівність не відхиляється.

Якщо порівнюються не дві, а декілька дисперсій, то замість F-критерію Фішера застосовують критерій Кокрена (G). Для k експериментів з однаковим числом вимірювань N маємо:

$$G = D_{\max}/(D_1 + D_2 + \dots + D_k). \quad (5.4)$$

Відношення максимальної із дисперсій до суми усіх інших має свій ймовірнісний розподіл, який залежить від статистичного рівня значущості, числа дисперсій і числа ступенів вільності $f = N - 1$. Теоретичні значини цього розподілу (G_T) табульовані.

Якщо значина G, розрахована згідно з (5.4), є менша за табличну (G_T), то усі k дисперсій вважаються однорідними на прийнятому статистичному рівні значущості. Коли ж $G > G_T$, то неоднорідна дисперсія $D_{\max}$ виключається, а із решти (k-1) перевіряється наступна максимальна. Порівняння здійснюють до тих пір, поки не отримують умову $G < G_T$.

Нехай, для прикладу, треба оцінити значущість наступних трьох дисперсій: $D_1 = 25,46 \text{ см}^2$; $D_2 = 36,96 \text{ см}^2$ і $D_3 = 15,16 \text{ см}^2$. Усі вони отримані у 3-х експериментах (k = 3) із числом вимірювань у кожному N = 10. Максимальною із приведених дисперсій є друга, тобто $D_{\max} = 36,96 \text{ см}^2$.

Із формули (5.4) отримуємо, що

$$G = D_{\max}/(D_1 + D_2 + D_3) = 36,96/(25,46 + 36,96 + 20,16) = 0,447.$$

Числа незалежних вимірювань при цьому є такими:

$$f_1 = k - 1 = 3 - 1 = 2;$$

$$f_2 = N = 10.$$

Згідно із табличними даними для $f_1 = 3$ і $f_2 = 10$ для довірчої імовірності 95% $G_T = 0,445$. Оскільки в даному випадку $G > G_T$, то оцінювана максимальна дисперсія D_2 є неоднорідною із іншими.

Оцінювання дисперсій D_1 і D_3 на однорідність здійснюємо за таких вихідних даних:

$$f_1 = k - 1 = 2 - 1 = 1;$$

$$f_2 = N = 10;$$

$$G = D_{\max}/(D_1 + D_3) = 25,46/(25,46 + 20,16) = 0,558.$$

Для $f_1 = 1$ і $f_2 = 10$ $G_T = 0,602$. Оскільки це більше за $G = 0,558$, то дисперсії D_1 і D_3 є однорідними і на статистичному рівні 0,05 різниця між ними є випадковою. Обидві вони, на відміну від дисперсії D_2 , представляють одну і ту ж генеральну сукупність.

5.2.3 Коефіцієнт варіації

Даний коефіцієнт (V) – це стандарт, виражений у відсотках до середньої арифметичної того чи іншого випадкового процесу:

$$V = (\sigma \cdot 100) / X. \quad (5.5)$$

Використання коефіцієнта варіації має сенс при вивченні варіювання параметра, який приймає тільки **позитивні** значини. Немає суті у величини V для процесу, середня значина якого дорівнює нулю.

Більше того, навіть для вибірок виключно з позитивними даними коефіцієнт варіації може не мати фізичного смислу. Для прикладу розглянемо процес коливання поздовжнього профілю поля. На практиці його реєстрацію здійснюють з допомогою профілографу. Основою цього приладу є рейка, яка встановлюється на певній висоті паралельно поверхні поля.

Припустимо, що у першому варіанті вимірювання ця рейка була розташована на висоті 0,5 м, а в другому – на висоті 0,3 м. В обох випадках відстань від рейки до поверхні поля фіксуватимемо в одних і тих же точках.

Середнє квадратичне відхилення (або стандарт) процесу, як відомо, не залежить від середньої арифметичної. В результаті для двох варіантів вимірювання профілю поля матимемо теоретично рівні значини σ . Але оскільки дійсні середні арифметичні при цьому у нас є різними, то різними, всупереч логіці, будуть і значини коефіцієнту варіації V .

Слід підкреслити, що для такого процесу, як вертикальні коливання профілю поля (шляху, дорого тощо) фізичного смислу не має і середнє арифметичне, так як висоту установки базової горизонтальної лінії, від якої здійснюються вимірювання, є результатом суб'єктивного вибору.

Мінливість (варіабельність) процесу прийнято вважати незначною, якщо коефіцієнт варіації, визначений за формулою (5.5), не перевищує 10%. При значині цього показника більше 10%, але менше 20% мінливість є середньою. Якщо $V > 20\%$, то мінливість випадкового процесу вважають значною.

Важливо, що при вивченні варіабельності параметрів однакової розмірності слід проявляти певну обережність. Коефіцієнт варіації може дати невірну оцінку щодо мінливості при різних середніх арифметичних, але однакових стандартах. В цьому випадку мінливість доцільніше оцінювати саме величиною середнього квадратичного відхилення.

5.2.4 Довірчий інтервал і довірча ймовірність

Оцінювання **точності** отриманого експериментального результату здійснюють з допомогою такої статистичної оцінки, як довірчий інтервал (D_i). Для його розрахунку використовують наступну формулу:

$$D_{\text{інт}} = X \pm t \cdot S_x,$$

де X – середня арифметична;
 t – теоретична значина критерію Стьюдента;
 S_x – похибка середньої арифметичної.

Границі $X - t \cdot S_x$ і $X + t \cdot S_x$ прийнято називати **довірчими**, а та ймовірність, з якою істинна (невідомо нам) значина параметру не виходить за межі цих границь, – **довірча ймовірність** (P).

Досить часто замість довірчої ймовірності P використовують величину, яка доповнює її до одиниці. Ця величина називається **статистичним рівнем значущості** і позначається символом α :

$$\alpha = 1 - P.$$

На практиці точність отриманого результату оцінюють при розв'язанні двох задач.

Перша із них така. Нехай здійснено $N = 120$ вимірювань глибини суцільної культиваци зябу і визначено її середнє арифметичне ($X = 10,4$ см) та похибку середнього ($S_x = 0,2$ см). Для заданої довірчої ймовірності $P = 95\%$ треба знайти відповідний довірчий інтервал $D_{\text{інт}}$.

Для цього з урахуванням P і N із таблиць знаходимо значину t -критерію. Наразі вона є такою: $t = 1,98$. Помноживши її на величину S_x , встановлюємо половину довжини довірчого інтервалу $t \cdot S_x = 1,98 \cdot 0,2 = 0,4$ см.

Звідси довірчий інтервал, за межі якого з ймовірністю P не вийде істинна значина вимірюваного параметру, становитиме:

$$D_{\text{інт}} = X \pm t \cdot S_x = 10,4 \pm 0,4 \text{ см.}$$

Друга задача формулюється так: по заданому розміру $\pm t \cdot S_x$ довірчого інтервалу встановити довірчу ймовірність P . Для цього спочатку знаходять розрахункову значину t -критерію (t_p):

$$t_p = t \cdot S_x / S_x.$$

Задавши $\pm t \cdot S_x = \pm 0,6$ см, знаходимо, що

$$t_p = t \cdot S_x / S_x = 0,6 / 0,2 = 3.$$

Після цього із таблиць з урахуванням t_p і N визначаємо, що цьому відповідає довірча ймовірність $P = 99,9\%$.

5.3. Лінійна та криволінійна кореляції

В наукових дослідженнях часто мають місце співвідношення між перемінними, коли кожній значині признаку X відповідає не одне, а кілька можливих значин признаку Y . Такі зв'язки, на відміну від функціональних, є стохастичними (імовірностними) або кореляційними. Вони характеризуються тісністю і формою зв'язку. Для встановлення останніх використовують спеціальні статистичні методи, які називаються кореляцією і регресією.

За формою кореляція може бути лінійною і криволінійною, а за напрямом – прямою і зворотною.

Під регресією розуміють зміну результативної ознаки Y (функції) за певної визначеної зміни аргументу – ознаки X . Зв'язок між функцією і аргументом виражається рівнянням регресії або кореляційним рівнянням.

Для оцінки тісноти зв'язку між функцією і аргументом застосовують коефіцієнт кореляції або кореляційне відношення.

Лінійною (прямолінійною) кореляційною залежністю між ознаками X і Y називають таку, яка описується рівнянням прямої лінії

$$Y = a + b \cdot X, \quad (5.6)$$

де a , b – сталі величини.

Якщо при збільшенні X величина Y зростає, то кореляція і регресія позитивні або прямі. В протилежному випадку – від’ємні або зворотні.

За числовий показник лінійного зв’язку використовують коефіцієнт кореляції – r . Він є безрозмірною величиною, яка змінюється в таких межах:

$$-1 < r < +1.$$

Причому, чим ближча значина r до $+1$ чи -1 , тим тісніший прямолінійний кореляційний зв’язок і навпаки. Коли ж $r = 0$, то між ознаками X і Y лінійний зв’язок відсутній (зате криволінійний може існувати!).

За сучасного рівня розвитку математичного програмного забезпечення (MathCAD, MathLAB, Statistica, Mathematica, Maple тощо) визначення сталих величин a і b лінійної регресії (5.6), а також коефіцієнта кореляції r та його стандартної похибки і критерію істотності є процедурою абсолютно формальною, тому окремо на цьому питанні зупинятися не будемо.

Дуже часто у науковців з’являється впевненість, що величина коефіцієнта кореляції на рівні 0,5 є досить високою і такою, за якої збіг варіації двох ознак (X і Y) має бути у половині всіх випадків.

Водночас, теорія кореляції свідчить, що ступінь взаємовпливу у варіюванні двох величин більш точно визначається квадратом коефіцієнта варіації (r^2). Величину r^2 називають коефіцієнтом детермінації і позначають R^2 . Цей коефіцієнт показує долю тих змін, які у вивчаємому явищі (процесі) залежать від змінюваного фактора (аргументу).

Наприклад, при $r = 0,5$ тільки 25% (а не 50%) зміни функції обумовлене зміною аргументу. У зв’язку з цим коефіцієнт детермінації є більш інформативним по відношенню до коефіцієнта кореляції. Коли, для прикладу, стверджують, що «90% вертикальних коливань трактора обмовлені коливаннями профілю поля», то тут 90% – це коефіцієнт детермінації, а не кореляції.

Якщо $r < 0,3$, то лінійну кореляційну залежність вважають слабкою, коли $r = 0,3 \dots 0,7$ – середньою, при $r > 0,7$ – сильною.

На відміну від прямолінійної, криволінійна регресія – це така залежність, коли за однакового приросту аргументу приріст функції є неоднаковим. Показником криволінійної залежності є кореляційне відношення (η). За незначного масиву дослідних даних (до 30) ця величина розраховується наступним чином:

$$\eta = \sqrt{\frac{\sum(Y - y_c)^2 - \sum(Y - y_x)^2}{\sum(Y - y_c)^2}},$$

де $\sum(Y - y_c)^2$ – сума квадратів відхилень індивідуальних значин Y від загальної середньої арифметичної y_c ;

$\sum(Y - y_x)^2$ – сума квадратів відхилень варіант від частинних середніх y_x , які відповідають фіксованим значинам незалежної ознаки Y .

Для розрахунку кореляційного відношення за цим виразом здійснюють наступні підготовчі операції. Спочатку значини незалежної ознаки Y ранжують за зростанням і розраховують середнє арифметичне u_c . Потім цей ряд чисел розбивають на $k = 4 \dots 7$ груп з таким розрахунком, щоб у кожній із них було не менше двох значин аргументу. Далі для кожної із k груп розраховують частинне середнє u_x , а вже потім – кореляційне відношення η .

За умови функціональної залежності Y від X кореляційне відношення дорівнює 1. Якщо $\eta = 0$, то це вказує на некорельованість ознак X і Y . За проміжного характеру кореляційної залежності кореляційне відношення знаходиться в межах

$$0 < \eta < 1.$$

Величину η^2 називають індексом детермінації. Він, по аналогії з коефіцієнтом кореляції, визначає відсоток варіювання ознаки Y (функції) від впливу аргументу X .

Похибку (S_η) і критерій істотності (t_η) кореляційного відношення η знаходять із наступних виразів:

$$S_\eta = \sqrt{(1 - \eta^2)/(N - 2)};$$

$$t_\eta = \eta/S_\eta,$$

де N – об'єм масиву даних, за якими розраховується значина η .

Теоретичну значину критерію t_η для того чи іншого статистичного рівня значущості можна знайти в таблицях багатьох посібників.

На практиці науковці часто необґрунтовано апроксимують експериментальні дані лінійною або нелінійною залежністю. В дійсності слід поступати так. Коли є підтвердження, що між ознаками X і Y існує прямолінійний зв'язок – слід застосовувати механізм прямолінійної кореляції.

Якщо заздалегідь відомо, що між ознаками X і Y криволінійний зв'язок, то з допомогою того чи іншого програмного забезпечення підбирають таку криволінійну залежність, яка найбільш точно відтворює зв'язок між функцією та аргументом досліджуваного процесу.

Водночас, досить часто характер зв'язку (лінійний чи криволінійний) між ознаками X і Y невідомий. В цьому випадку, як би криволінійно не розташовувались експериментальні значини функції, слід діяти за наступним алгоритмом. Спочатку розраховують коефіцієнт лінійної кореляції r і кореляційне відношення η . Далі визначають дійсну значину F -критерію:

$$F_d = \frac{(\eta^2 - r^2) \cdot (N - k)}{(1 - \eta^2) \cdot (k - 2)}.$$

Якщо виявиться, що дійсна значина F -критерію менша за табличну (тобто $F_d < F_T$), то зв'язок між ознаками X і Y є лінійним, а коли $F_d \geq F_T$ – криволінійним.

5.4. Методика аналізу кривих за відсутності оптимуму

Результати виробничих або економічних процесів зазвичай залежать від багатьох факторів. Якщо ми зафіксуємо всі із них, окрім одного, то будемо мати

залежність, яка описується функцією однієї змінної. Найчастіше представляють інтерес точки, в яких описана функція приймає своє найменше або найбільше значення. Для знаходження таких точок екстремуму використовують диференціальне числення. Далі будемо розглядати процеси, в яких область зміни фактора можна умовно розбити на два інтервали. В кожному з інтервалів залежність результату від досліджуваного фактора є близькою до лінійної, але на одному з інтервалів функція зростає або спадає швидко, а на іншому – повільно. Опуклі криві, які описують такі функції, будемо називати **двозонними**.

Прикладами таких двозонних кривих є графіки функцій

$$y_1 = \frac{\alpha x}{1 - \beta x^2} \text{ на проміжку } (0, \beta^{-0.5})$$

та

$$y_2 = k - \alpha e^{-\beta x} \text{ на проміжку } (-\infty, \infty).$$

Типові графіки цих функцій зображені на рис. 5.4. Функція y_1 , для прикладу, застосовується для апроксимування кривих буксування мобільних енергетичних засобів. Функція y_2 використовується при аналізі низки процесів з експоненціальною складовою.

При дослідженні процесів, які описуються двозонними кривими, бажано вміти визначати точку, яка розділяє проміжки зміни фактора з істотно різними швидкостями залежності результату від цього фактору. Для цього потрібно дати строге означення точки (будемо називати її точкою оптимуму двозонної кривої), яка поділяє двозонну криву на дві частини з суттєво різними властивостями та запропонувати метод визначення її положення.

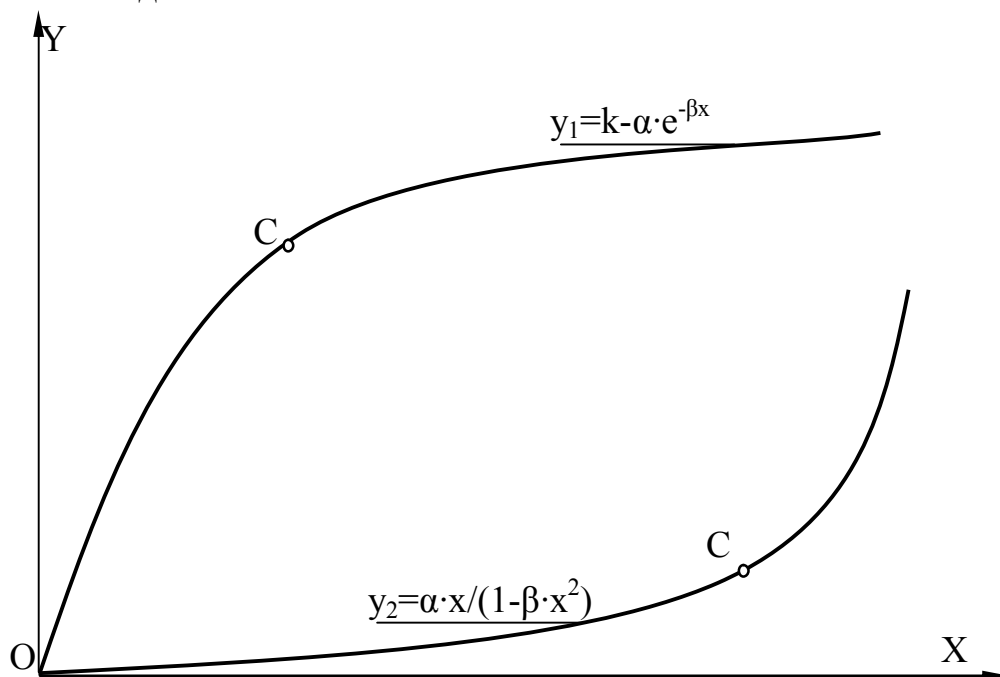


Рис. 5.4 – Типові графіки двозонних кривих

Для означення точки оптимуму розглянемо «ідеальну» двозонну криву, яка представляє собою дволанкову ламану. Очевидно, що для неї точкою оптимуму буде

кутова точка. Для довільної двозонної кривої розглянемо множину дволанкових ламаних, вписаних в цю криву. Тобто таких ламаних, початок і кінець яких співпадає з початком і кінцем цієї кривої, а кутова точка лежить на цій кривій.

Точкою оптимуму двозонної кривої будемо називати кутову точку дволанкової ламаної, яка найкращим чином наближає задану криву.

Нам потрібно визначити критерій, який визначає ступень наближення. В якості такого критерію візьмемо відстань між функціями $f(x)$ та $g(x)$, які описують криву та ламану відповідно, в просторі $L[a, b]$:

$$\rho(f, g) = \int_a^b |f(x) - g(x)| dx. \quad (5.7)$$

Іншими словами, ламана тим краще наближає криву, чим менша площа фігури, обмеженої цими лініями.

Безпосереднє використання формули (5.7) для отримання аналітичного виразу для координат точки оптимуму ускладнене, оскільки наявність модулів не дозволяє скористатися апаратом диференціального числення. Але використання інформації про опуклість функції дає можливість спростити міркування.

Для прикладу розглянемо графік функції $f(x)$, яка на заданому кінцевому проміжку $[a, b]$ є опуклою вгору (рис. 5.5).

Розглянемо кінцеві точки кривої $A(a, f(a))$ та $B(b, f(b))$. Позначимо кутову точку ламаної через C . Вона буде мати координати $C(c, f(c))$. Із викладеного вище випливає, що потрібно так обрати точку оптимуму C , щоб площа S фігури, обмеженою кривою ACB та ламаною ACB була мінімальна.

Зрозуміло, що для фіксованої кривої величина S буде залежати тільки від положення точки C , тобто $S = S(c)$. В силу опуклості кривої ACB матимемо, що ламана лежить під кривою, а, отже S дорівнює різниці площ, обмежених кривою та ламаною.

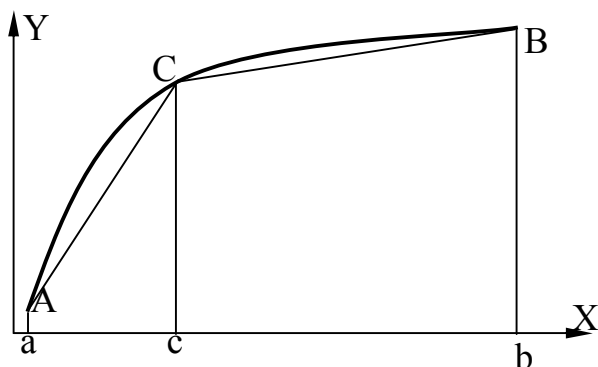


Рис. 5.5 – Ламана, вписана в криву

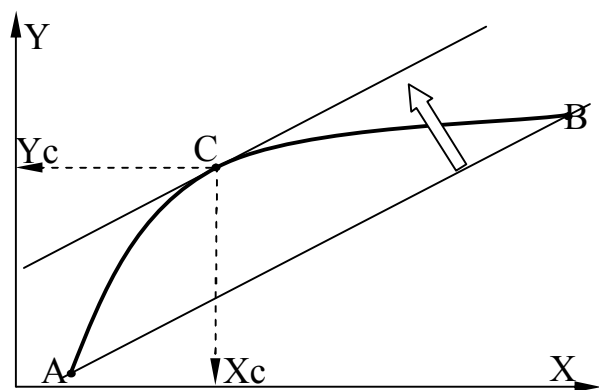


Рис. 5.6 – Графічний метод пошуку точки оптимуму

Площа обмежена кривою, є постійною величиною. Таким чином S буде тим менше, чим більшою буде площа $P(c)$ під ламаною. Отже умову $S(c) \rightarrow \min$ можна замінити еквівалентною умовою

$$P(c) \rightarrow \max. \quad (5.8)$$

Вираз для площі $P(c)$ можна записати явно, оскільки розглянута фігура складається з двох трапецій:

$$P(c) = \frac{f(a) + f(c)}{2}(c - a) + \frac{f(c) + f(b)}{2}(b - c). \quad (5.9)$$

Для знаходження максимуму $P(c)$ розв'яжемо рівняння $P'(c) = 0$. Оскільки

$$2P'(c) = f'(c)(b - a) + f(a) - f(b),$$

то абсциса C точки оптимуму C визначається із умови

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \quad (5.10)$$

Зауважимо, що для диференційованих функцій точка, визначена умовою (5.10) обов'язково існує в силу теореми Лагранжа. Майже ті ж самі міркування будуть правильними, якщо будемо розглядати функцію, опуклу вниз. Тільки нам буде потрібно шукати не максимум, а мінімум функції $P(c)$. Умова (5.10), як необхідна умова екстремуму, буде мати місце і в цьому випадку.

Графічна інтерпретація умови (5.10) полягає в наступному. Точка C є оптимумом в означеному сенсі тоді і тільки тоді, коли дотична до неї паралельна відрізку AB (рис. 5.6). Це дає можливість досить просто визначати координати шуканої точки C безпосередньо з графіка функції. Для цього слід взяти лінійку, прикласти її до точок A і B і рухати паралельно уверх до тих пір, поки над нею не залишиться єдина точка дуги ACB (рис. 5.6).

Цілком зрозуміло, що вираз (5.10) можна використовувати для аналітичного визначення положення (координат) точки C . В якості прикладу розглянемо функцію

$$y = \frac{\alpha x}{1 - \beta x^2}. \quad (5.11)$$

На рисунку 5.7 зображено її графік на відрізку $[a, b] = [0, 0.5]$ при наступних значеннях параметрів: $\alpha = 0.12$, $\beta = -3$. На кінцях відрізка вона приймає наступні значення: $y(a) = y(0) = 0$, $y(b) = y(0.5) = 0.24$.

Оскільки для функції (5.11) похідна, в загальному вигляді, дорівнює

$$y'(x) = \alpha \frac{1 + \beta x^2}{(1 - \beta x^2)^2},$$

то умова (5.10) при заданих значеннях параметрів представляє собою рівняння

$$0.12 \frac{1 + 3c^2}{(1 - 3c^2)^2} = \frac{0.24}{0.5}.$$

Ми отримали бікватратне рівняння, яке на розглянутому проміжку $[0, 0.5]$ має єдиний корінь $c = 0.37$. На рис. 5.15 показано, що отриманий результат узгоджується з встановленою геометричною інтерпретацією, тобто його можна отримати шляхом зсуву прямої, яка з'єднує кінці кривої.

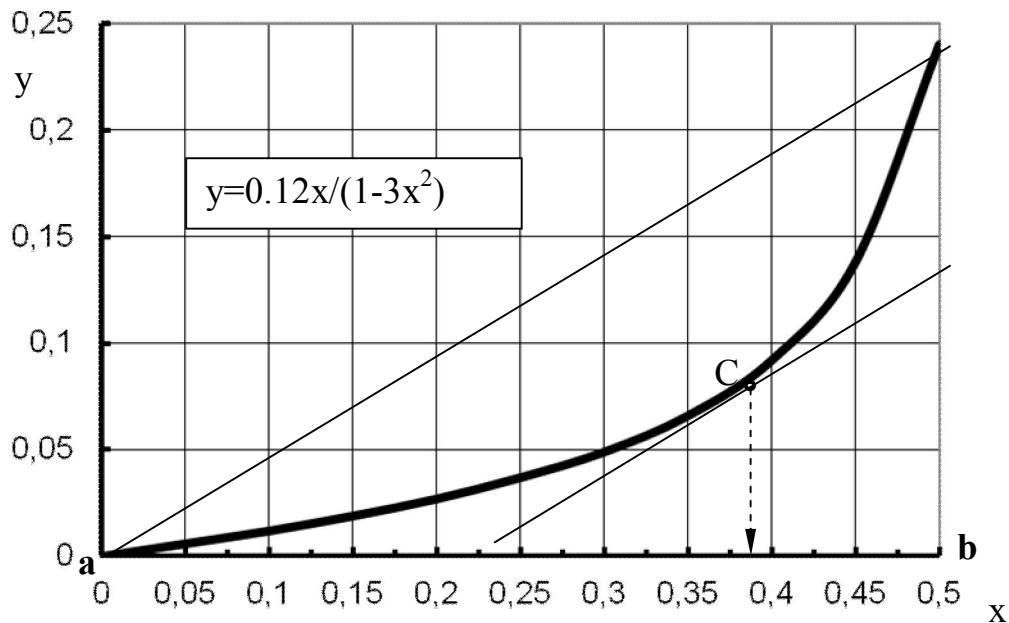


Рис. 5.7 – Практичне визначення координат т. С

Питання для самоперевірки

1. Як використовувати критерій Ірвіна для перевірки дослідних даних на «сумнівність»?
2. Розкажіть суть методики остаціонарювання масиву дослідних даних за середньою значиною.
3. Як перевірити нуль-гіпотезу про рівність/нерівність двох середніх значин двох випадкових процесів?
4. У чому полягає суть методики застосування найменшої істотної різниці для порівняння двох середніх значин випадкових процесів?
5. У чому полягає некоректність застосування похибки вибіркової середньої значини для установлення точності досліду?
6. Як порівнювати між собою дві дисперсії двох випадкових процесів?
7. Розкажіть про особливості методики порівняння трьох і більше дисперсій між собою.
8. Для яких випадкових процесів коефіцієнт варіації не носить фізичного смислу, а тому не може застосовуватися при аналізі?
9. Для чого досліднику потрібні довірчий інтервал і довірна ймовірність?
10. У чому полягає суть методики визначення характеру зв'язку між функцією і аргументом: лінійна чи нелінійна?
11. Що таке індекс детермінації?
12. Суть методики аналізу кривих за відсутності у них оптимуму.